DV-PredNet: Biologically Plausible Video Next Frame Prediction with Higher-level Semantics

Anonymous submission

Paper ID 16

Abstract

This paper investigates biologically plausible video next-frame prediction in the domain of high-frequency physical interactions. We explore the limitations of PredNet, a deep network implementing predictive coding, on a custom dataset designed to isolate the spatiotemporal behaviors of dynamic objects. To address these limitations, we introduce DV-PredNet (Dorsal+Ventral PredNet), a disentangled, two-stream architecture to separately model physical dynamics ('where') and visual appearance ('what'). Our model demonstrates improvements in both visual fidelity and trajectory tracking. However, we identify a characteristic performance degradation during high-impact events, such as collisions. Here, the model prioritizes learned visual statistics over enforcing physical consistency, resulting in a persistent one-frame lag. This reactive behavior reveals a fundamental limitation of the predictive coding framework with purely implicit physics learning, pointing towards the need for stronger physical priors or hybrid architectures to achieve physically reliable dynamics.

1. Introduction

Machine learning and artificial intelligence have long drawn inspiration from the brain's underlying neuroscience principles. Biologically plausible systems offer insight into higher-level cognitive capabilities and general intelligence, both of which possess potential to substantially improve performance and robustness of current deep learning architectures [5]. A key aspect of human cognition is 'intuitive physics,' referencing our capacity to make accurate inferences about the physical world by running approximate mental simulations [2]. A central goal in machine learning is to build models functionally similar to said intuitive physics engine (IPE), enabling them to understand and predict the dynamics of their environment.

A common approach towards this goal is to create physics-informed models that explicitly incorporate the physical priors. However, despite said models' remarkable progression in controlled environments, they often struggle to generalize in stochastic, real-world situations, a challenge known as the 'sim-to-real' gap [11]. Given the limitations of physical priors, this work investigates whether a model with no explicit physics priors can learn the underlying rules of a dynamic system. To borrow from our cognitive system, we implement predictive coding, a biologically plausible computational framework explored in deep learning models [4, 13]; this theory posits that the brain is constantly predicting incoming sensory input and updating its beliefs based on the prediction's accuracy [10].

As a first step toward learning in stochastic environments, we pose a more focused question: can a predictive coding model implicitly learn the rules of a controlled, deterministic, high-frequency physical setting? To test this, we first analyze the performance of PredNet, a prominent implementation of predictive coding that has demonstrated impressive performance in next-frame prediction, particularly in the domain of autonomous driving [9]. However, in these low-frequency scenarios, minor inaccuracies have less impact on the qualitative perception of the prediction, meaning pixel-level reconstruction losses can prove to be effective. We hypothesize that this approach will be less physically robust in videos involving low visual complexity and high-frequency dynamics, such as rapidly colliding objects.

To test this hypothesis, we designed a simple synthetic dataset involving the interaction of two bouncing balls in an enclosed space. In this work, we first investigate the limitations of the baseline PredNet on this dataset. We then propose DV-PredNet, a novel, disentangled two-stream architecture inspired by humans' visual processing stream, where a 'what' pathway is responsible for object representation and a 'where' pathway is responsible for spatiotemporal dynamics, mirroring the ventral and dorsal pathways in the striate cortex respectively [3].

While our modifications exhibit substantial improvements, we also identify a noticeable degradation in performance during moments of dynamic interaction.

2. Methods

Our model builds upon the PredNet architecture [9], a deep recurrent network inspired by the predictive coding inference principle. PredNet is a hierarchy of recurrent modules (ConvLSTMs), where each layer generates a top-down prediction of the activity in the layer below it. A bottom-up error signal, representing the discrepancy between the prediction and actual activity, is used to update the recurrent states. We adapt this error-propagation mechanism for our two-stream model.

2.1. Disentangled Two-Stream Architecture

Inspired by the human visual pathway, the core principle of our model involves a disentanglement of object representation ('what') and spatiotemporal dynamics ('where'). These two parallel pathways are built from the original PredNet components as defined in equations (1) to (4) in the original paper [9]. These pathways are jointly trained, although each is tasked with learning a specialized objective. To create a clear hierarchy for our learning objectives, we define the red ball as the primary object of interest. The final frame prediction, $\hat{x}_t$, is generated by a rendering head that convolves the concatenated outputs from both the what-stream (R_0^t) and the where-stream (D_0^t) to synthesize the final image (see Eq. 1).

To ensure both streams have access to valuable contextual information (i.e. environmental features like walls), we introduce a shared contextual pathway. While canonical models propose that only error signals are passed up bottom-up [1], we found that providing a direct sensory stream stabilized learning, especially considering the where-stream lacks contextual information otherwise. A lightweight encoder processes the input frame at each timestep which is concatenated with the error signal in the bottom-up pass. This provides both pathways with an abstract representation of the scene.

2.2. What Pathway

The 'what' pathway is responsible for rendering a visually accurate prediction of the scene based on the outputs of its recurrent modules.

To move beyond blurry, pixel-level predictions and focus on higher-level semantics, we introduce a perceptual loss [7]. This is done by extracting feature maps from the penultimate convolutional layer of a pretrained VGG16 model; a perceptual loss is computed by measuring L1 distance between the feature maps of the predicted and ground-truth frames.

Additionally, we compute a weighted Mean Absolute Error (MAE) loss. A segmentation mask of the salient object (the red ball) is extracted from the ground truth image; a MAE loss is computed across the entire image, but the loss in the salient masked region is amplified by a scalar of $\gamma = 15$. While pixel-wise losses have been shown to produce blurry images [6, 12], it has also been shown to act as a valuable stabilizer for adversarial or perceptual training objectives [6]. The addition of the mask encourages the model to prioritize high-fidelity construction of the object rather than allocating unnecessary attention to the background.

2.3. Where Pathway

The 'where' pathway serves as our model's implicit physics engine. Its objective is to learn an intuitive understanding of physical interactions and causality without being explicitly programmed with physics equations.

The primary objective of the where-stream is to predict the segmentation mask of the primary object at the next time step. The loss is a combination of Dice and Focal loss, which are robust for segmentation tasks and prioritize the accurate prediction of the foreground objects over the static background [8].

A linear coordinate head is attached to the where-stream's output state and is trained to predict a 6-tuple, $\hat{c}_t$, which represents the normalized 3D coordinates of both balls in world space (see Eq. 2 and Eq. 3). Predicting the coordinates of both objects rather than just the primary ball of interest encourages the model to develop a higher-level representation of their interaction and resulting dynamics.

$$\hat{x}_t = \sigma\left(\text{Conv}\left(\left[R_0^t; D_0^t\right]\right)\right) \tag{1}$$

$$s_t = \text{ReLU}(\text{Linear}(\text{Flatten}(\text{MaxPool}(\text{ReLU}(\text{Conv}(D_0^t)))))) \tag{2}$$

$$\hat{c}_t = \text{Linear}(s_t) \tag{3}$$

2.4. Loss Function

The model is trained end-to-end by minimizing a composite loss function. The total loss, $\mathcal{L}_{\text{total}}$, is a weighted sum of components for state prediction (location $\mathcal{L}_{\text{loc}}$, mask prediction $\mathcal{L}_{\text{mask}}$), visual reconstruction (perceptual $\mathcal{L}_{\text{vgg}}$, weighted MAE $\mathcal{L}_{\text{mae}}$), and internal error minimization for each stream ($\mathcal{L}_{\text{err}}$, $\mathcal{L}_{\text{err\_loc}}$):

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_{\text{loc}}\mathcal{L}_{\text{loc}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{mae}}\mathcal{L}_{\text{mae}} \\ + \lambda_{\text{vgg}}\mathcal{L}_{\text{vgg}} + \lambda_{\text{err}}\mathcal{L}_{\text{err}} + \lambda_{\text{err\_loc}}\mathcal{L}_{\text{err\_loc}}\end{aligned} \tag{1}$$

The weighting coefficients were empirically determined through hyperparameter tuning to be: $\lambda_{\text{loc}} = 0.2$, $\lambda_{\text{mask}} = 0.7$, $\lambda_{\text{err\_loc}} = 0.1$, $\lambda_{\text{mae}} = 0.4$, $\lambda_{\text{vgg}} = 0.5$, $\lambda_{\text{err}} = 0.1$.

Furthermore, we introduce event-based loss weighting. In moments of collision, the loss contributions for all relevant metrics for the current and subsequent timestep are amplified by a scalar factor of $\gamma = 15$ (determined by a sign-flip in ground-truth velocity vectors). This forces the model to prioritize learning from these high-impact, physically salient events.

3. Experiments

3.1. Custom Dataset

We generated a synthetic dataset in Unity to test intuitive physics learning; in each scene, two balls, one red and one of random color, were initialized in an enclosed box with a random velocity and position.The primary object was designated a red color to ensure a clean and consistent segmentation mask extraction via HSV thresholding. This mask was then dilated to emphasize the object's boundary, ensuring that the object is contained within the mask and is not constructed outside of it. This approach intentionally simplifies the segmentation task to provide a near-perfect, noise-free ground truth signal for the 'where' stream. By removing segmentation error, our experiments are tasked to focus exclusively on evaluating the core contribution of this work: the effectiveness of the disentangled predictive architecture, particularly the 'where' stream, in learning high-frequency dynamics. To ensure consistent physical behavior, all rigidbodies were assigned a dynamic friction coefficient of 0.4, a static friction coefficient of 0.6, and a bounciness (elasticity) coefficient of 0.9. Variability in other factors (e.g. texture, lighting) were eliminated. The dataset consists of 300 simulations (128x128, 150 frames each) which was processed with a sliding window and randomly partitioned into 7549 training and 200 validation clips of 15 frames each. Ground-truth position and velocity vectors of both balls were recorded for each frame.

3.2. Training Details

The model was trained for 80 epochs using the Adam optimizer (initial LR 10^{-3}, decaying to 10^{-4} in the latter half of training) with a batch size of 16 and 'clipnorm' of 1.0. Experiments were conducted on a single NVIDIA A100 GPU via Google Colab. Each 80-epoch training run took approximately one to two hours to complete. The first two timesteps of each sequence were omitted from the loss calculation as a meaningful physics-based prediction requires at least two initial frames to infer trajectory and velocity.

4. Results & Discussion

We evaluated our model against the baseline on a challenging sequence involving multiple collisions and a brief occlusion. The baseline PredNet suffered from several qualitative issues: blurriness, shape inconsistency, vanishing objects, position inaccuracies, and poor semantic understanding of the scene. This can be attributed to over-reliance on L1 pixel loss, which leads to blurry images and phases out high-frequency dynamic details in favor of reconstructing the static background [6], as well as a poor understanding of spatiotemporality.

Our augmented DV-PredNet, however, demonstrates quantitative improvement in understanding of the scene's physics. A low validation location loss of 0.0148 (MAE) indicates the model is accurately tracking object trajectories. This is further supported by superior performance on standard benchmarks (SSIM, PSNR) and a significantly lower perceptual loss (see Table 1).

Model	Perceptual Loss ↓	SSIM ↑	PSNR ↑
PredNet	0.596	0.979	34.14
DV-PredNet	**0.138**	**0.988**	**38.59**
DV-PredNet($\mathcal{L}_{\text{Centroid}}$)	0.159	0.981	35.01

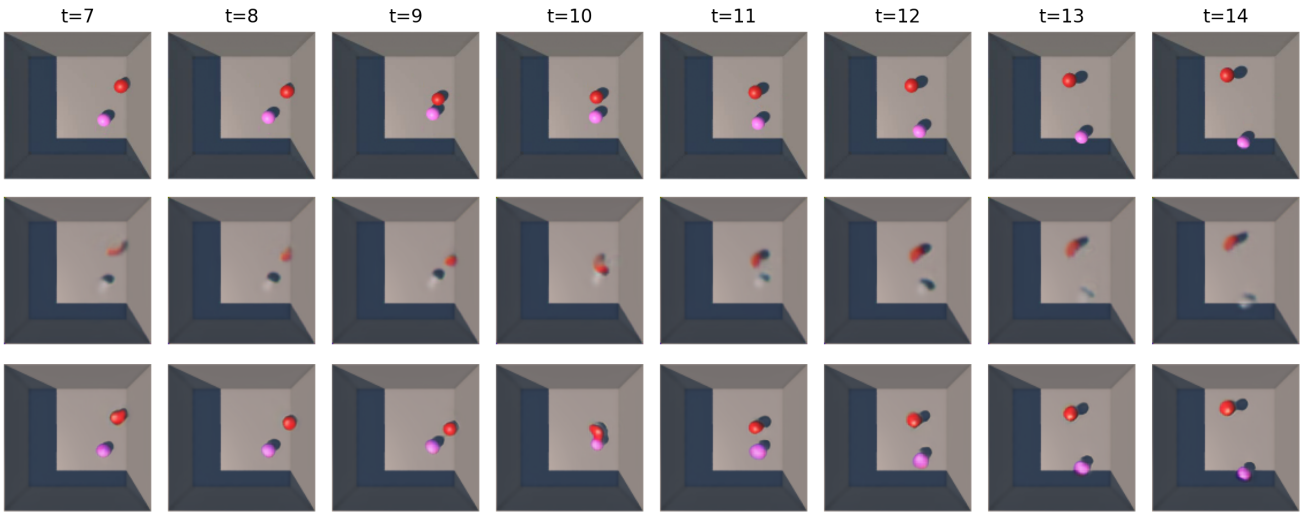
Table 1. Quantitative comparison of baseline PredNet vs. our proposed DV-PredNet. DV-PredNet shows significant improvements across all metrics. Lower is better for Perceptual Loss; higher is better for SSIM and PSNR.

Qualitatively, the predictions exhibit high fidelity reconstruction of both objects. However, as shown in Figure 1b, which displays the predicted segmentation masks, the model's performance degrades at key physical events. At $t = 9$, the red ball bounces off the wall, and at $t = 10$, it collides with the other ball. In both instances, the model fails to anticipate the sharp change in trajectory, resulting in a noticeable one-frame lag. The prediction at the time of collision reflects the pre-collision trajectory, and the model only corrects its course in the subsequent frame. This suggests that while the model has learned the dynamics, its predictive mechanism is more corrective than anticipatory.

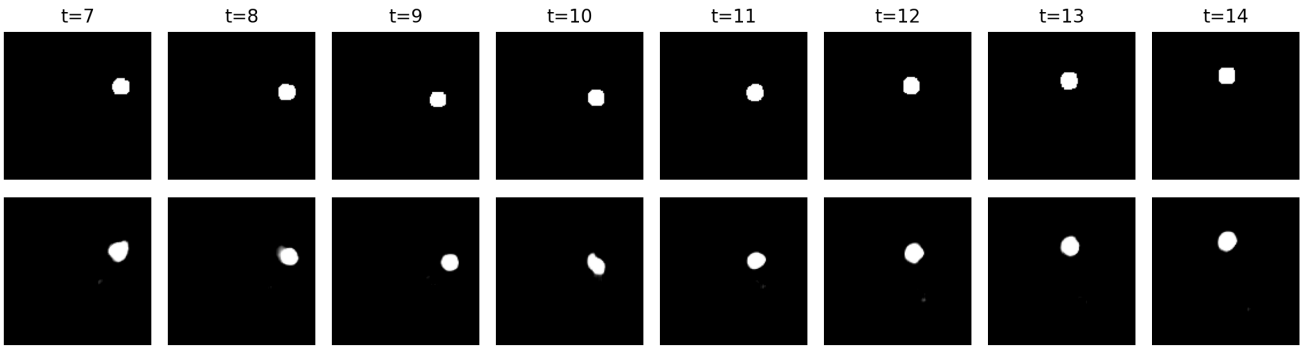
4.1. Ablation Study

An ablation study was conducted to weigh the contribution of newly introduced components (see Table 2). As expected, ablating the 'what' stream and its perceptual objectives resulted in a catastrophic degradation across all metrics, confirming its essential role in rendering a visually coherent scene. Similarly, removing the shared context encoder also harmed performance, indicating that providing a direct, abstract representation of the scene is critical for stabilizing the learning of both pathways.

Surprisingly, however, the variant without the 'where' stream's objectives achieved a slight improvement in quantitative metrics, with a lower Perceptual Loss (0.129 vs. 0.138) and a higher PSNR (39.13 vs. 38.59). This counterintuitive result indicates that the spatial signal from the 'where' stream, while essential for physical accuracy (as shown in Figure 2a), can act as a subtly conflicting prior that slightly degrades the final render quality. Taken together, these ablations confirm that while all components contribute meaningfully, the primary challenge lies not in the architectural disentanglement itself, but in designing specialized objectives that are fully compatible and do not create counter-productive interference.



(a) Qualitative comparison of predictions (t=7 to t=14). Top: Ground Truth. Middle: Baseline. Bottom: DV-PredNet.



(b) Predicted segmentation masks from DV-PredNet (t=7 to t=14), showing a one-frame lag at collisions (t=9, t=10).

Figure 1. Overall visual results. (a) shows DV-PredNet's significant improvement in object fidelity. (b) reveals the model's reactive lag during collisions.

Variant	Perceptual Loss ↓	SSIM ↑	PSNR ↑
DV-PredNet	0.138	0.988	38.59
No 'Where' Stream	**0.129**	**0.988**	**39.13**
No 'What Stream'	1.34	0.742	15.16
No Context Encoder	0.210	0.982	35.51

Table 2. Ablation study results. The "No 'Where' Stream" variant shows a slight improvement in visual fidelity metrics, highlighting an objective misalignment challenge.

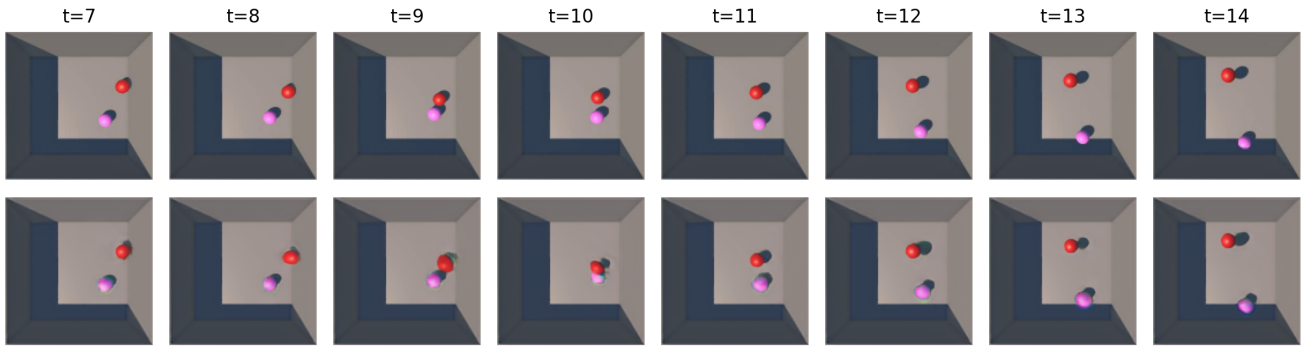
4.2. Physically Plausible Prediction

To further investigate the interplay between the 'what' and 'where' streams, we analyzed a model configuration trained with a heavily-weighted centroid loss and an amplified weight for the masked prediction loss ($\lambda_{\text{mask}} = 1.1$). This configuration resulted in worse quantitative performance (see Table 1) and a noticeable degradation in visual fidelity, producing rendering artifacts near the pink ball and worse shape accuracy; however, it was the only configuration out of extensive experimentation to successfully anticipate the collision and produce a physically plausible rebound, overcoming the one-frame lag present in our quantitatively superior models.

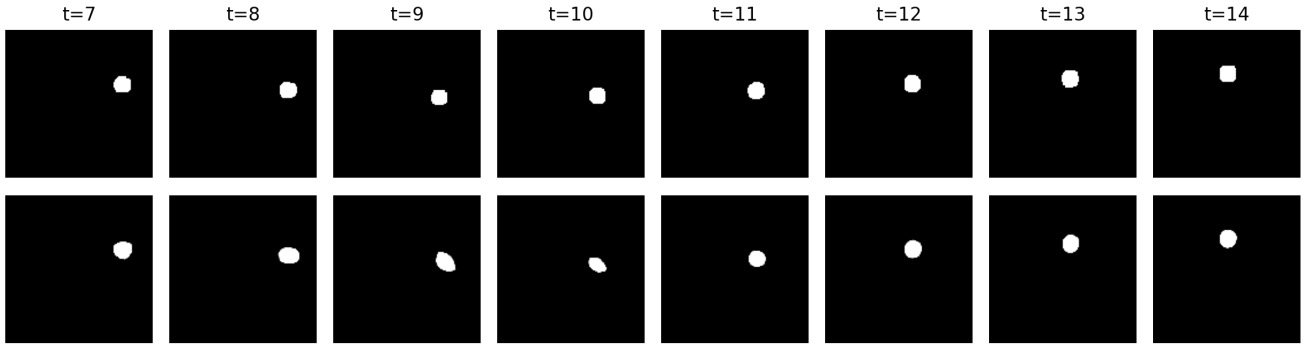
This finding provides strong evidence for an inherent trade-off between visual fidelity and physical dynamic robustness in our model. The aggressive, location-based objective provided the necessary corrective "pull" to overcome the reactive tendencies of the predictive coding model, but did so at the expense of visual coherence.

5. Conclusion

In this work, we investigated the performance of a biologically-inspired, predictive coding model for learning intuitive physics in a deterministic environment. We introduced a disentangled, two-stream model that separates



(a) Qualitative comparison of DV-PredNet with heavy focus on 'where' stream (t=7 to t=14). Top: Ground Truth. Bottom: DV-PredNet with centroid loss and amplified λ_{mask} ($\lambda_{\text{mask}} = 1.1$).



(b) Predicted segmentation masks from DV-PredNet with heavy focus on 'where' stream (t=7 to t=14)

Figure 2. Overall visual results. In subfigure (a) at t=9, the red ball is visibly detached from the wall, indicating a successful rebound off the wall, though visual fidelity is reduced. Similarly, the mask at t=9 in subfigure (b) is more closely aligned vertically to the GT mask compared to the prediction of DV-PredNet in 1a.

learning appearance and dynamics. Despite achieving substantial improvements in object fidelity and location tracking, our central finding was the identification of a key limitation: a reactive one-frame lag during non-linear collisions.

Our analysis further revealed a fundamental trade-off between visual fidelity and physical accuracy in our model's implementation. A model configuration with a heavily-weighted physical objective was the only one to produce an anticipatory, physically plausible rebound, but at the cost of significant visual degradation. This conflict was verified by ablation studies where removing the physical objective paradoxically improved visual metrics. These findings provide a clear case study demonstrating that purely implicit, data-driven approaches can struggle to learn true causal dynamics.

Our work argues that the future of physically reliable world models lies in a "middle ground." Purely implicit physical models may prove to be insufficient in outputting a consistent, robust understanding of dynamics and causality, and therefore motivating hybrid, physics-guided architectures that integrate learned representations with explicit physical priors. However, there still remains potential in purely implicit models, and future works can further explore deeper avenues of disentanglement for robust physical reasoning without explicit priors.

References

[1] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

[2] Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. Available at: https://www.pnas.org/doi/10.1073/pnas.1306572110.

[3] M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1): 20–25, 1992.

[4] K. Han, H. Wen, Y. Zhang, D. Fu, E. Culurciello, and Z. Liu. Deep predictive coding network with local recurrent processing for object recognition, 2018. arXiv preprint arXiv:1805.07526.

[5] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick.

Neuroscience-inspired artificial intelligence. *Neuron*, 95(2): 245–258, 2017.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks, 2018. arXiv preprint arXiv:1611.07004.

[7] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. arXiv preprint arXiv:1603.08155.

[8] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[9] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning, 2017. arXiv preprint arXiv:1605.08104.

[10] R. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999.

[11] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world, 2017.

[12] X. Wang, S. López-Tapia, A. Lucas, X. Wu, R. Molina, and A. K. Katsaggelos. A general method to incorporate spatial information into loss functions for gan-based super-resolution models, 2024. arXiv preprint arXiv:2403.10589.

[13] H. Wen, K. Han, J. Shi, Y. Zhang, E. Culurciello, and Z. Liu. Deep predictive coding network for object recognition, 2018. arXiv preprint arXiv:1802.04762.